

<https://doi.org/10.1038/s41524-024-01264-z>

Complexity of many-body interactions in transition metals via machine-learned force fields from the TM23 data set

Check for updates

Cameron J. Owen^{1,2,5}✉, Steven B. Torrisi^{2,4,5}, Yu Xie², Simon Batzner², Kyle Bystrom², Jennifer Coulter², Albert Musaelian², Lixin Sun² & Boris Kozinsky^{2,3}✉

This work examines challenges associated with the accuracy of machine-learned force fields (MLFFs) for bulk solid and liquid phases of *d*-block elements. In exhaustive detail, we contrast the performance of force, energy, and stress predictions across the transition metals for two leading MLFF models: a kernel-based atomic cluster expansion method implemented using sparse Gaussian processes (FLARE), and an equivariant message-passing neural network (NequIP). Early transition metals present higher relative errors and are more difficult to learn relative to late platinum- and coinage-group elements, and this trend persists across model architectures. Trends in complexity of interatomic interactions for different metals are revealed via comparison of the performance of representations with different many-body order and angular resolution. Using arguments based on perturbation theory on the occupied and unoccupied *d* states near the Fermi level, we determine that the large, sharp *d* density of states both above and below the Fermi level in early transition metals leads to a more complex, harder-to-learn potential energy surface for these metals. Increasing the fictitious electronic temperature (smearing) modifies the angular sensitivity of forces and makes the early transition metal forces easier to learn. This work illustrates challenges in capturing intricate properties of metallic bonding with current leading MLFFs and provides a reference data set for transition metals, aimed at benchmarking the accuracy and improving the development of emerging machine-learned approximations.

Molecular dynamics (MD) simulations can reveal atomistic mechanisms for a wide range of fundamental material, chemical, and biological processes. Ab initio methods like density functional theory (DFT) can calculate atomic forces, energies, and stresses, but are too expensive for MD simulations at large time- and length-scales. Approximate surrogate models referred to as interatomic potentials or ‘classical’ force fields (FFs) have bridged this gap but can take months or even years to develop, where practitioners exhaustively fit FF parameters to properties like experimental quantities (e.g., melting point, bulk lattice constant, structure factors, etc.)¹. While their fixed analytical forms make them efficient and interpretable^{2–7}, predictions from classical FFs are limited in transferability beyond their initial training targets even for the same chemical system. Application to complex

chemistries and phenomena like bond-breaking in reactive systems requires great care and close supervision, as e.g. assumptions that go into capturing bonded interactions can make decisive differences in simulation outcomes^{8,9}.

In response to these challenges, atomistic FF development has, over the past decade, been revolutionized by the advent of machine-learned force fields (MLFFs), where the FF construction task is reduced to fitting a surrogate model of flexible form to first-principles data. Simple analytical forms of traditional FFs are replaced with flexible universal approximators to achieve increased accuracy and transferability^{10,11}. This has allowed practitioners to fit MLFFs on demand for any desired system that can be computed by ab initio methods. This approach has already yielded successes for

¹Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA. ²John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. ³Robert Bosch LLC Research and Technology Center, Watertown, MA 02472, USA. ⁴Present address: Toyota Research Institute, Los Altos, CA, USA. ⁵These authors contributed equally: Cameron J. Owen, Steven B. Torrisi.

✉ e-mail: cowen@g.harvard.edu; bkoz@g.harvard.edu

simple single-element systems, where MLFFs have been used to reveal surprising long-range mechanical behavior using MD^{12,13} and yield highly accurate and expressive power for determining a variety of material properties.

Whether studying materials systems of one element or many, reliable reference data sets are of incredible importance to the task of MLFF model training and benchmarking^{14–16}, where model architectures can be compared using the same set of training and test labels. Presently, no dedicated benchmark data set exists for the *d*-block of the periodic table, which makes it difficult to compare model architectures across a common standard in this set of elements, important for a wide range of applications e.g. heterogeneous catalysis on bulk and nanoparticle (NP) structures, (high-entropy) alloys, surface reconstructions, and metallurgy. Another major challenge, which reference benchmark databases can help address, is the need to broker a compromise between efficiency and accuracy in the choice of ML formalism and hyperparameters, which also depends on the complexity of the system to be modeled. Simplifications in the representation fidelity of atomic environments or in model architecture typically come at a cost to accuracy of the predictions made by the resulting MLFF, which must be weighed against the demands of the target application. To further complicate this task, users typically lack the means and data to gauge what level of model architecture and representation fidelity is needed when approaching new systems, i.e., it is hard to know in advance when a simpler model will suffice.

Benchmark data sets and subsequent studies have been curated before within the community for solid-state materials: previous work carefully benchmarked the performance of a wide variety of model forms (GAP, MTP, NNP, SNAP, and qSNAP), with Zuo et al.¹⁷ releasing the associated data set. This data set focused on a variety of structural motifs across a set of six elements (Li, Mo, Cu, Ni, Si, and Ge) chosen to represent different electronic character (metallic vs. covalent/semiconducting).

The models trained in this work, and in a followup investigation in ref. 18, highlight larger force and energy errors on early transition metals, like Mo, as opposed to later transition metals, like Cu. By explicitly comparing these metallic systems across model architectures in terms of predictive accuracy on energies and forces, hints of the strong disparity in model performance appear, but were not commented on in further detail. More comprehensive reviews of the performance of model architectures like GAP and MTP on body-centered-cubic transition metals and alloys have been performed^{19,20}, but the discussion of predictive accuracy was, again, not at the forefront. Despite the potential of high predictive errors, correlated with where the models are trained on the periodic table, these efforts have demonstrated that one can still obtain proper physical descriptions of the systems (e.g. phonons, alloy compositions, etc.), but there has been no systematic investigation into these model performances across the transition metals, which would yield increased understanding of the problem elements, and push the field towards better model architectures.

Another common benchmark data set used to compare accuracy of MLFFs is MD17^{21–24} which is comprised of small, organic molecules in vacuum containing main-group elements (i.e., C, O, N, and H), where bond topology is typically rigid and many models can achieve chemical accuracy (< 1 kcal/mol)¹¹. While the latter is useful for the molecular chemistry community, a reference data set for transition metals would provide enormous benefit for the heterogeneous catalysis and metallurgical communities, among others tasked with building MLFFs for these elements.

Thus, better understanding of the tradeoffs between efficiency and efficacy in leading MLFFs for targeted elements could help to drive new methods development, as well as accelerate future model training. Moreover, there are increasingly many options for practitioners: MLFF development is now well into its second decade of application^{25–28}, and many improvements have been made to the fitting processes, with uncertainty-based active learning^{10,29–31} followed by exact mapping onto low-dimensional surrogate models (e.g., polynomial and spline models)^{29,32,33}. A plethora of MLFF architectures exist: such as MTP³⁴, GAP³⁵, ACE³⁶, MACE³⁷, PAiNN³⁸ SNAP³⁹, SchNet⁴⁰, DeepMD⁴¹, with each model

architecture exhibiting its own strengths and weaknesses. More recently, equivariant neural network methods, e.g. NequIP and Allegro^{11,42} have been shown to accurately predict the behavior of a diverse range of molecular and materials systems ranging from solid-state ionic diffusion and heterogeneous catalysis to small molecules and water. The models considered in this work (FLARE and NequIP) represent two of the recent leading MLFF approaches with inherent differences in how the representations are constructed for atomic environments.

Consideration of inherently more challenging material and chemical systems, however, will inevitably prompt further MLFF development. To this point in time, MLFFs have demonstrated near-chemical accuracy on available organic molecule benchmark data, but have shown mixed results on materials¹⁷. By extending the composition and structural space wherein these MLFFs operate, novel model architectures that achieve state-of-the-art accuracy on organic systems may have to evolve from their current form to accomplish this task. In hinting to the results presented here, we find that high angular resolution of NequIP is required to improve the predictive accuracy for more difficult transition metals in TM23, albeit at significantly increased computational cost. Hence, novel model architectures are required that combine efficiency with representation resolutions to accurately capture these difficult solid state systems.

Such a data set curation task has the potential to benefit more than just MLFF developers. Within the domains of surface science, heterogeneous catalysis, and alloys, crucial mechanisms such as surface restructuring⁴³, active site dynamics, and cooperative reaction mechanisms on metal surfaces occur on long time- and length-scales⁹, making gains in MLFF efficiency and accuracy critical for the interpretation of experimental observables. With important use-cases like this in mind, it behooves us to understand the role that varying atomic representations and model fidelity can play across machine-learning architectures, and to make the TM23 data set available for the MLFF development and broader scientific community. More available testing benchmarks will aid the overall project of developing and judging both purpose-built and general purpose FFs which can be flexibly applied to a wide variety of systems.

Finally, of central importance to this work is a demonstration of a new mode of use of MLFFs as a probe to extract fundamental physical and chemical trends from first-principles reference data, and to obtain a better understanding of the relationship of the predictive errors to the parameters employed in the underlying quantum mechanical method. Specifically, by varying the DFT and MLFF model hyperparameters, such as the electronic temperature, or body-order and angular resolution of the representation, and comparing the accuracy with respect to first principles calculations, we can directly assess the complexity of the quantum many-body interatomic interactions and connect it to domain-knowledge intuition in terms of electronic structure of different metals.

Results and discussion

Overview of the benchmarking task

We begin by presenting the TM23 data set, comprised of ab initio molecular dynamics simulations of 27 *d*-block metals, from which we sample a subset of training structures and associated energy, force, and stress labels from high-fidelity DFT calculations. These benchmark data are then used to evaluate the performance of two different MLFF architectures employing different representations of the atomic environments. To limit the breadth of comparison between the swath of available MLFF approaches, we only focus on two architectures: (1) Gaussian processes based on the atomic cluster expansion descriptors and (2) equivariant neural networks. Open-source implementations of these architectures are employed, namely FLARE¹⁰ and NequIP¹¹, respectively. By evaluating a large set of relevant elements across two model architectures, it is demonstrated that even chemically simple systems – mono-elemental bulk materials with a single vacancy in low-temperature crystalline, high-temperature crystalline, and molten phases – present significant challenges for accurate learning by these leading models and a persistent trend of errors across the *d*-block of the periodic table.

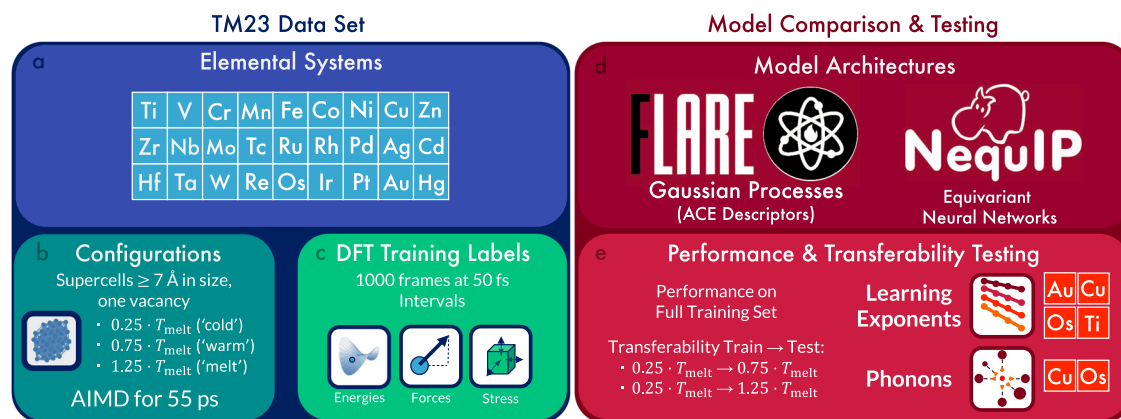


Fig. 1 | TM23 data set construction and model testing overview. **a** Complete table of transition metals studied. **b** AIMD sampling of atomic environments and the size requirement for super-cell creation. **c** Extraction procedure and high-fidelity DFT calculations. **d** Models were trained on high-fidelity labels via the FLARE and

NequIP codes. **e** We gauged the model performance by using test sets from all temperatures, and by assessing transferability from 'cold' to 'warm' and 'cold' to 'melt' sets of data. We also compute learning exponents for Au, Cu, Ti, and Os and phonon dispersion curves for Cu and Os.

In benchmarking the accuracy of these methods, trade-offs that arise with differing model fidelities are explored and reasonable references in accuracy are provided for other practitioners approaching these systems. A schematic of the benchmarking workflow is provided in Fig. 1, and the explicit details of each component are provided in the "Methods" Section. Importantly, clear trends in accuracy are found across the *d*-block which hold regardless of model architecture, as quantified by relative errors of force, energy, and stress. These benchmarks demonstrate the role that model architectures play across the same set of first principles reference data, where the choice of model parameters can influence their performance in a marked fashion for the more 'difficult' metals as opposed to 'easier' metals. The TM23 data set is made freely available (see "Data Availability" Section) in order to facilitate direct comparison across the wide range of actively developed MLFF methods, and it is envisioned that this will act as a useful benchmark comparison target within the computational materials science community.

Lastly, we note that the data set used here was not intended to provide high-fidelity models for dynamic evolution of these systems. Rather, the primary task was to uncover the drastic differences in predictive accuracy across transition metals as well as the fundamental relationships between the DFT parameters employed, meaning that future work will consider the dynamic evolution of these systems and, importantly, if further augmentation of the data set is necessary.

TM23 Data Set Description

Super-cells for 27 transition metals were created from their experimentally verified ground state crystal structures, as provided by the Materials Project⁶ at their 0 K lattice constants predicted using the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional⁴⁴. Super-cells were generated with a requirement that each lattice vector was at least 7 \AA in length. This value was chosen to balance the number of unique atomic environments observed within the finite radius cutoff of MLFF local structure representations against the number of atoms required for DFT calculation. A single vacancy was introduced into each super-cell in order to diversify non-trivial atomistic configurations. Ab initio molecular dynamics (AIMD) simulations were then performed at three temperatures: $0.25 \cdot T_{\text{melt}}$, $0.75 \cdot T_{\text{melt}}$, and $1.25 \cdot T_{\text{melt}}$, where T_{melt} is the experimental melting temperature, for a total of 55 ps. Each system was evaluated at a single volume as defined by the ground state lattice parameters. We henceforth refer to these temperatures as 'cold', 'warm', and 'melt', respectively. Melting temperatures for each system were extracted from ref. 45. We note that even though these are short time-scale AIMD simulations, the corresponding radial distribution functions for each temperature were analyzed to confirm the loss of crystalline-order at increased temperature. Examples of these

crystalline and amorphous radial distribution functions are provided in Supplementary Figs. 57–83. This analysis is provided as a Jupyter notebook paired with the data set on Materials Cloud.

The first 5 ps of each trajectory was used for thermal equilibration at the desired temperature, and subsequently, representative frames (structure snapshots) were extracted from the remaining 50 ps of the trajectory at 50 fs intervals, in order to reduce their correlation. Extracted frames were used as input for high-fidelity static DFT calculations, which produced the energy, force and stress labels used for model training and testing. To accelerate each AIMD trajectory, *k*-point sampling was limited to only the Γ -point, whereas finer *k*-point grids were used for generating training labels for each extracted frame. Spin-polarization was not included for all high-fidelity DFT frames, which is discussed in more detail in the "Methods" Section. This procedure yielded 1000 frames for each of the three AIMD temperatures and thus 3000 in total for each metal. The entire collection of high-fidelity frames (81,000 in total) will be provided via the Materials Cloud upon publication.

MLFF training on TM23

We then trained and tested both FLARE and NequIP models using energy, force and stress labels for each metal using the set of 3000 high-fidelity static DFT calculations. For our first test, we obtained a holistic overview of model performance by using as wide a training/testing set as possible. For each element, the training set drew from a combined set of 2700 frames using the first 900 extracted frames of each AIMD trajectory at the three temperatures. The validation set for NequIP was chosen to be 10% of the training set, selected randomly. The training weights in NequIP for forces and stress were set to 1, while the energies also employed a coefficient of 1, but used the 'PerAtomMSELoss' as discussed in the NequIP GitHub repository. In the FLARE code, weights were not set for the energy, forces, and stresses, rather the noise hyperparameters were initially set according to the discussion in the "Methods" Section, and are optimized over the course of model training. The test set contained a total of 300 frames taken to be the final 100 frames from each AIMD trajectory at the three temperatures considered. Since prediction accuracy is influenced by the model parameters, e.g., representation cutoff, radial and angular bases for FLARE, or neural network depth and angular resolution in the case of NequIP, these model parameters were explicitly tested for each metal using a grid search on a smaller training set (200 frames) and test set (100 frames). Results for each model architecture employing the best parameters are shown here in the main text, and the complete set of values is provided in Supplementary Note 6. To put these results in the context of model size, we also provide the descriptor dimensions used for the FLARE models and number of weights used in the NequIP models in Table 1. The descriptor dimension, n_d is calculated using Eqn. (1),

Table 1 | Training statistics for FLARE and NequIP

Elem.	FLARE		NequIP		
	ACE Dim.	Total Train. Time (hr)	N_{wts}	Train. Speed (hr/epoch)	GPU Arch.
Ag	180	18.799	265580	4.037	A100
Au	220	39.308	612744	3.680	A100
Cd	364	11.365	265580	3.375	A100
Co	330	13.747	265580	2.760	A100
Cr	270	9.904	265580	4.728	A100
Cu	140	16.363	265580	7.271	V100
Fe	462	10.499	612744	46.000	V100
Hf	225	8.348	265580	5.440	V100
Hg	225	9.869	265580	1.780	A100
Ir	765	18.367	612744	17.121	A100
Mn	462	13.497	265580	13.455	V100
Mo	330	11.357	265580	11.074	A100
Nb	225	10.711	265580	3.234	A100
Ni	330	18.375	612744	10.897	V100
Os	225	11.833	265580	1.961	A100
Pd	225	10.420	612744	25.407	V100
Pt	225	10.270	265580	1.486	A100
Re	225	14.124	265580	3.484	A100
Rh	330	13.770	265580	0.850	V100
Ru	225	12.128	612744	6.453	V100
Ta	225	11.543	265580	10.428	A100
Tc	225	15.695	265580	2.695	V100
Ti	270	16.619	265580	17.391	V100
V	270	12.413	265580	11.146	V100
W	330	11.342	265580	8.543	A100
Zn	364	13.429	265580	3.288	A100
Zr	225	8.138	612744	2.564	A100

FLARE ACE descriptor dimensions, total training times, and number of CPUs, as well as NequIP network weights, training speeds, and GPU architectures employed for the models used to generate the results in Fig. 2. All NequIP models were trained using a single GPU.

given as

$$n_d = (n_{\text{max}} \cdot n_{\text{species}} + 1) \cdot n_{\text{max}} \cdot \frac{n_{\text{species}}}{2} \cdot (\ell_{\text{max}} + 1). \quad (1)$$

We then provide interpretation of the underlying trends across compositions in predictive accuracy, and examine the influence of model error on observable material properties.

Accuracy comparison of FLARE and NequIP

Before proceeding, we comment upon the choice of error metric for each of the target labels. Force errors are expressed as percentages, defined in Equation (2), rather than mean absolute values (e.g., MAEs or RMSEs), since these quantities correlate with the average magnitude of the force and simultaneously the AIMD temperature, which naturally varies across metals.

$$\% \text{Error} = \frac{\text{MAE}}{\text{MAV}} \cdot 100 \quad (2)$$

Full MAEs for forces, energies, and stresses are provided in Supplementary Note 5. A comparison of force MAEs to the melting temperature are

provided in Section S2 as a demonstration of this relationship. Hence, test percent errors for the best FLARE and NequIP models on TM23 are provided in Fig. 2 where trends in accuracy across the d -block can be immediately observed. Moving from left to right across Fig. 2, we see that early TMs exhibit higher test errors across forces, energies, and stresses relative to the late Pt-group and coinage metals (Groups IX, X, and XI, respectively), with the coinage metals producing the lowest errors of the entire set.

A trend across models can also be observed, where NequIP test errors are systematically lower than FLARE using ACE B2 descriptors. FLARE test errors obtained using the ACE B1 descriptor are provided in the Supplementary Information, which are systematically higher than FLARE using ACE B2 and NequIP. The observed trend between FLARE B2 and B1 ACE descriptors can be explained by an increase in effective body-order, where the 2-body B1 descriptor with a kernel power of 2 yields an effective but not complete 3-body interaction between environments, while the 3-body B2 descriptor at the same power yields 5-body terms. Despite an overall improvement in the force % errors using NequIP compared to either of the FLARE models, the same trends remain persistent when comparing model performance across the d -block metals, in that early transition metals exhibit noticeably higher errors than late transition metals across all model architectures. Additionally, we note that we explicitly explored the relationship between observed model accuracy and k -point density, as is provided in Supplementary Fig. 1. We employ the minimum k -point density as a label for the inherent accuracy of the DFT calculations, and find no correlation to model test error.

Curiously, when moving from the coinage metals in Group XI to the Group XII transition metals (Zn, Cd, and Hg), an increase in the % error is also observed. However, Group XII metals exhibit the lowest melting points of the elements considered in this study, meaning that the absolute magnitudes of the forces are small. Coupling this observation to the fact that the force MAEs for Group XII systems are on the order of only 10s of $\text{meV} \text{ \AA}^{-1}$, it can be reasoned that these systems are sampling near the inherent noise of DFT given the convergence protocols, rather than reflecting difficulties in model learning.

The % error values in Fig. 2 were determined using the entire training set of 2700 frames and testing on the remaining 300 frames. Corresponding mean absolute error values for the same models and data are provided in Fig. 3. The training procedures for each model architecture are described in explicit detail in the “Methods” Section, and differ since FLARE implements a Gaussian process whereas NequIP is a neural network. FLARE uses sparsification and MPI parallelization to circumvent the memory bottleneck present when considering the full training set of 2700 frames, where sparse representative atomic environments are selected from each training frame by the predictive uncertainty of the GP⁸. This yielded a total sparse set of 2700 atomic environments. On the other hand, the NequIP model trains using all atomic environments from each frame.

Transferability of FLARE and NequIP across temperatures and phases of TM23

Measuring a model’s ability to generalize beyond the training set distribution is of central interest for the development of MLFFs: a long simulation especially involving reactions and structural evolution may encounter new configurations not sampled in the training set. A valuable advantage of the TM23 data set is that the training data contains three distinct temperatures for each metal. This provides an opportunity to explore the extrapolative ability of ML models between temperatures and structural phases, where thermal disorder can produce dramatically different atomic environments. This latter statement is confirmed through the analysis of the radial distribution functions of the trajectories, where $0.25 \cdot T_{\text{melt}}$ retains crystalline order, reflected by the persistence of well-defined peaks which broaden and disappear at higher temperatures, indicating a transition to a molten phase.

We explore model transferability across temperatures in a similar fashion to an earlier work on motion of a single-molecule¹⁶. Models are thus trained using low temperature frames ($0.25 \cdot T_{\text{melt}}$) and are tested using frames from higher temperatures ($0.75 \cdot T_{\text{melt}}$ or $1.25 \cdot T_{\text{melt}}$). Results for

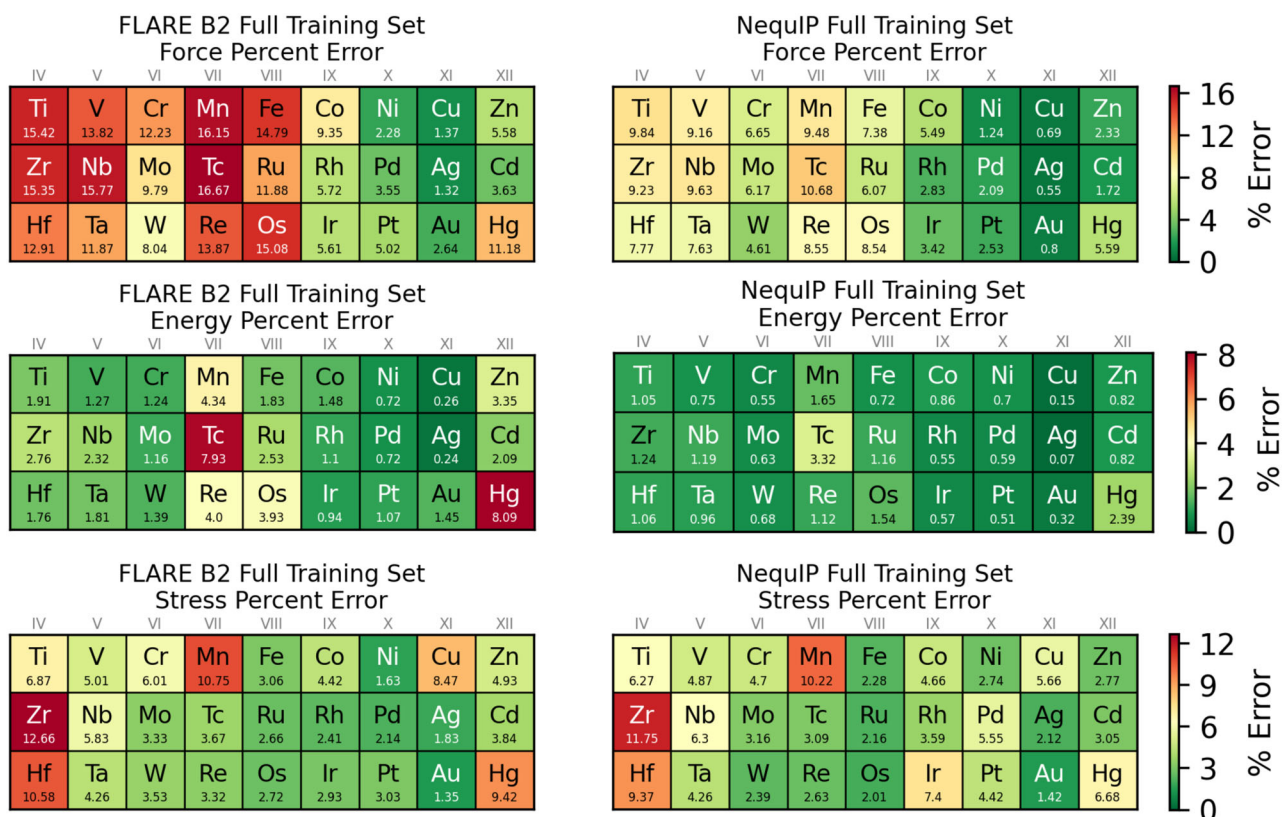


Fig. 2 | Full training and test set force, energy, and stress percent (%) errors for the 27 TM systems. Results are from FLARE, using the B2 invariant ACE descriptors with $\zeta = 2$ and NequIP. Errors for NequIP are almost uniformly lower than for FLARE (with some exceptions in stress prediction). A version of this figure containing the MAEs is provided in Fig. 3, and B1 results and MAEs for all other labels

and models are provided in the Supplementary Information. The layout for each panel reflects the d -block of the periodic table, starting from Group IV (Ti, Zr, & Hf) and extending to Group XII (Zn, Cd, & Hg). A uniform color scheme is employed across all models and elements, where green represents the lowest % errors whereas red represents the highest % errors.

FLARE (using ACE B2) and NequIP models are provided in Figs. 4 and 5. We note similar accuracy trends to the full multi-temperature training set results of Figs. 2 and 3. With regards to metal-dependent performance, Group IX, X, and XI metals yield markedly lower errors than early transition metals (Group VIII and below) in forces, energies, and stresses for both transferability tasks. Trivially, both FLARE and NequIP yield lower predictive errors when tested on ‘warm’ as opposed to the ‘melt’ frames. From the model architecture perspective, the global trend noted in the previous section remains consistent, namely that NequIP outperforms FLARE at the B2 descriptor, yielding lower errors across nearly all of the metals.

Disparate learning behaviors of NequIP across TM23

In an effort to better understand the error as a function of the available training set size and model parameters, we investigated the effect of the NequIP architecture on the training exponent of a subset of these systems, a procedure discussed in the original NequIP paper¹¹. It has been observed that the test error of deep learning systems follows a power law of the form $\epsilon = aN^b$, where ϵ refers to the predictive error, N is the training set size, and a and b are constants. In ref. 11, it was shown that equivariant interatomic potentials, with the tensor rank, or angular rotation order of spherical harmonics, $\ell_{max} \geq 1$ exhibit higher values of b as compared to invariant methods ($\ell_{max} = 0$), meaning they learn faster with the number of data points. Here, we increase the value of ℓ_{max} (see “Methods” section for more detail) within the NequIP architecture, and demonstrate differences in the learning behavior. The complete set of model weights, training speeds, irreducible feature coefficients, and computational architecture are provided in Table 2. To determine differences in the learning behavior across a subset of the metals, we vary ℓ_{max} from 0 (limiting the model to invariant scalar features) to $\ell_{max} = 3$ (a fully equivariant model). For two metals (Au and Os),

ℓ_{max} is increased further to 4 and 5 to explore the effect of even higher angular resolution of the equivariant representation. Models are trained on data sets ranging from 100–900 frames taken from the $1.25 \cdot T_{melt}$ AIMD trajectory for each metal, in increments of 200, and force MAEs are employed to study learning dependencies on model architecture.

Au and Cu are chosen as a representative subset of the ‘easy’ metals, whereas Ti and Os represent the more ‘difficult’ systems. From Fig. 6, we draw three immediate conclusions. The first is that the error magnitude for all metals is markedly affected by increasing ℓ_{max} . Secondly, the learning exponent m differentiates each model’s ability to learn the forces as more training data is made available on a metal-basis: slopes are larger for the Au and Cu models using $\ell_{max} = 3$, relative to Os and Ti, mean that Au and Cu models learn faster with new data than Os and Ti. The Cu and Ti models are truncated at $\ell_{max} = 3$ given the results for Au and Os, where increasing the angular resolution to higher values increases computational cost of inference but does not bring about marked increases in predictive accuracies.

Finally, a subtler and provoking conclusion from these data is that the observed increase in learning exponent, as a function of increasing ℓ_{max} , revises previous understanding of the learning dynamics for equivariant models. In the original work by¹¹, power-law exponents were computed for the water data set of Cheng et al.⁴⁷ It was observed that the absolute value of the learning exponent increased when moving from $\ell_{max} = 0$ to $\ell_{max} > 0$ and the magnitude of the test error decreased, meaning that equivariant models with $\ell_{max} > 0$ learn faster and attain lower overall error. That work also observed ‘diminishing returns’, as the absolute value of the learning exponent increases the most from $\ell_{max} = 0$ to $\ell_{max} = 1$, significant, but smaller increases from $\ell_{max} = 1$ to $\ell_{max} = 2$, and then successive increases in ℓ presenting more modest changes to the learning exponent. While we find that increasing ℓ_{max} tends to reduce the overall magnitude of error in all cases, we

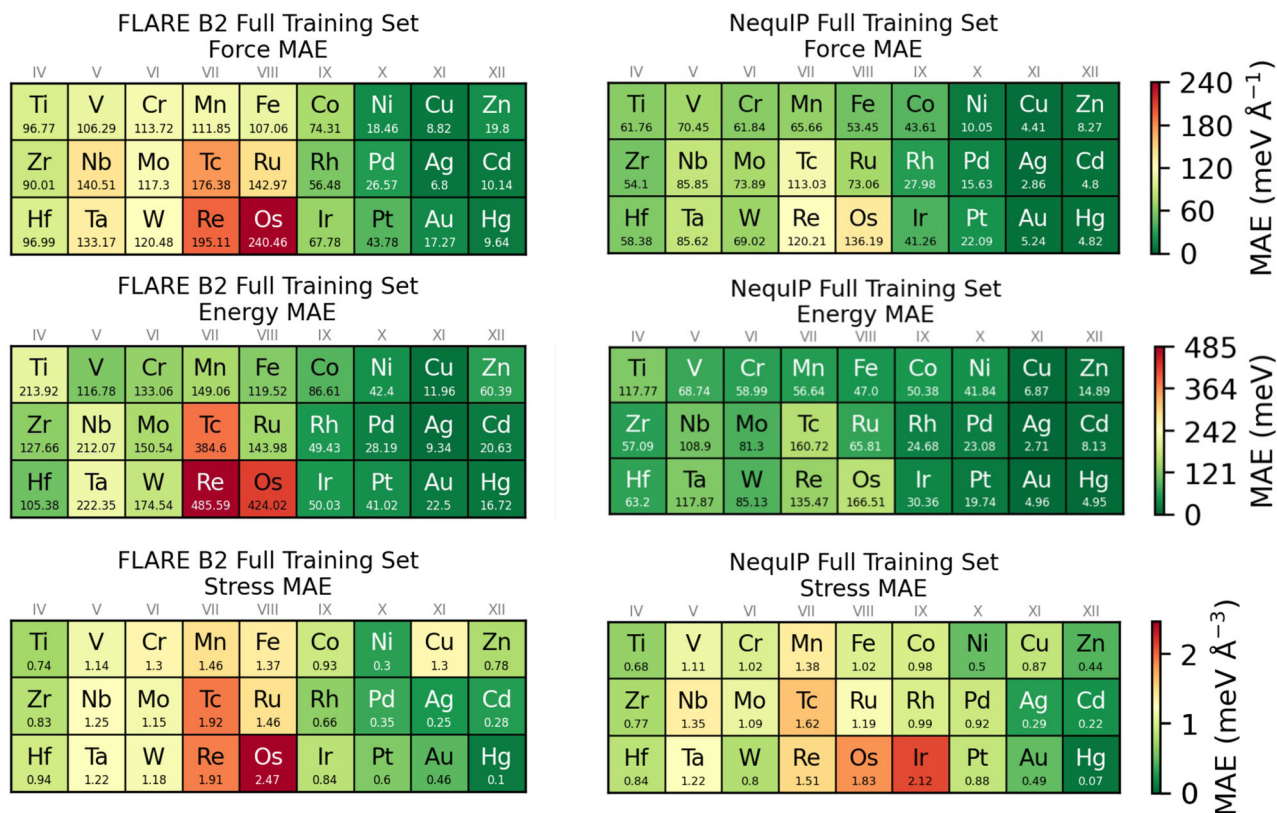


Fig. 3 | Full training and test set force, energy, and stress mean absolute errors (MAEs) for the 27 TM systems. Results are from FLARE, using the B2 invariant ACE descriptors with $\zeta = 2$ and NequIP. Aside from the difference in error reporting, this Figure is otherwise laid out identically to Fig. 2. Note that MAE label values scale with the magnitude of the forces, which in turn scale with the melting temperature

T_m of particular metals due to referencing AIMD temperatures against T_m on a per-metal basis- in other words, higher melting point metals tend to have higher MAEs not only due to model error but to larger variation in the force values. See Supplementary Note 2, Supplementary Fig. 2, and Supplementary Figs. 30-56 for further details.

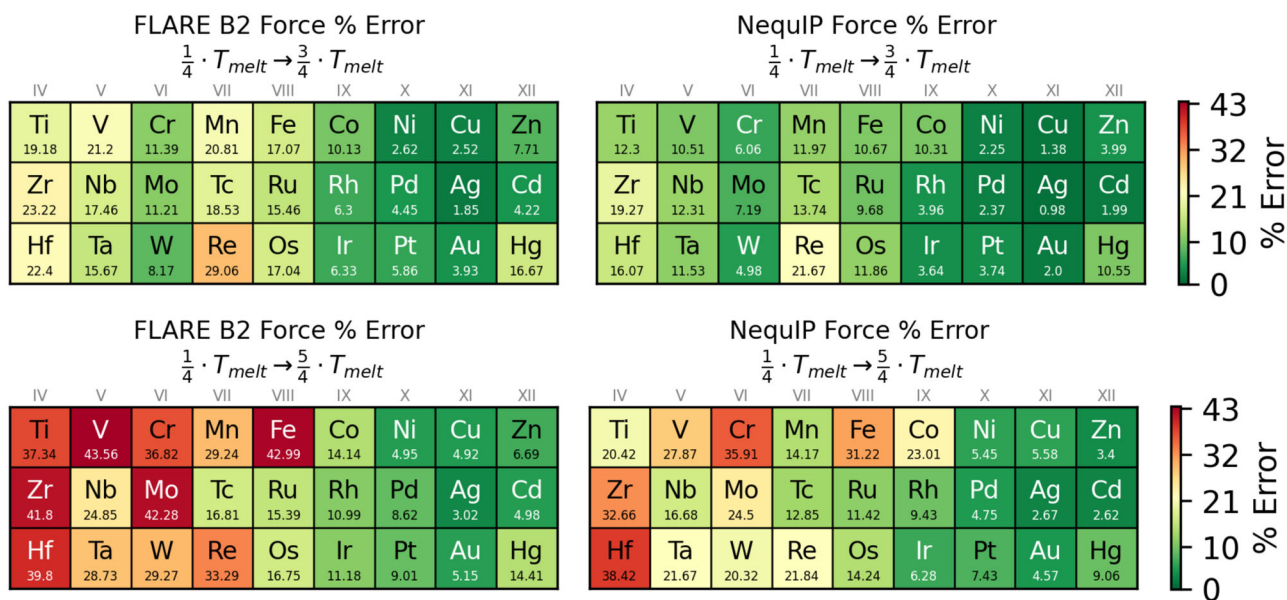


Fig. 4 | Force percent (%) errors for the 27 TM systems across temperatures. Training was done on 1000 frames at $0.25 \cdot T_{melt}$ and testing on either 1000 frames at $0.75 \cdot T_{melt}$ (left panels) or $1.25 \cdot T_{melt}$ (right panels). FLARE values are obtained

using the ACE B2 descriptor at $\zeta = 2$. The formats again reflect the d -block of the periodic table. A uniform color scheme is employed across all models, where green represents the lowest % errors whereas red represents the highest % errors.

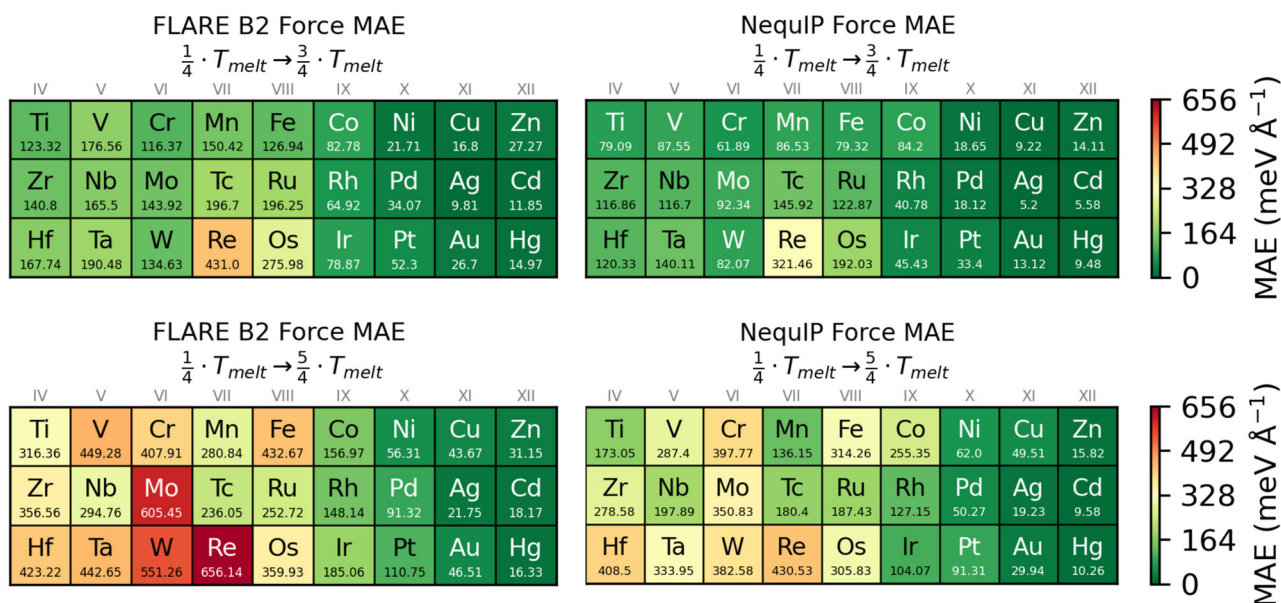


Fig. 5 | Force MAEs for the 27 TM systems across temperatures. Training was done on 1000 frames at $0.25 \cdot T_{melt}$ and testing on either 1000 frames at $0.75 \cdot T_{melt}$ (left panels) or $1.25 \cdot T_{melt}$ (right panels). Layout is otherwise identical to Fig. 4.

also find that the changes in learning rate differ across all metals. However, the higher angular-resolution models do still present an overall lower error, with evidence of possible saturation for high ℓ_{max} values. Furthermore, the biggest relative changes in the learning exponents do not occur between $\ell_{max} = 0$ to $\ell_{max} = 1$ in all four cases (Au, Cu, Os, and Ti), contrasting with the case for water in ref. 11, where the influence of equivariance was largest

from $\ell_{max} = 0$ to $\ell_{max} = 1$. Here, we see that the learning exponent steadily increases with increasing ℓ_{max} for Au until $\ell_{max} = 4$ and Ti until the upper bound of $\ell_{max} = 3$ is reached. For Cu, the effect of ℓ_{max} seemingly saturates between 2 and 3 for Cu, which is similar to Os. We note that the number of parameters (weights) and internal equivariant operations/tensor products in the neural network varies significantly with ℓ_{max} , which may also confound these results.

Table 2 | NequIP network weights, training speeds, and GPU architectures employed

Elem- ℓ_{max}	N_{wts}	Train. Speed (hr/epoch)	Comp. Arch.
Au ₀	76346	0.019	V100
Au ₁	137274	0.068	V100
Au ₂	142138	0.223	V100
Au ₃	134970	0.551	V100
Au ₄	122234	0.818	V100
Au ₅	106974	1.126	V100
Cu ₀	75832	0.029	A100
Cu ₁	136760	0.216	A100
Cu ₂	141624	0.911	A100
Cu ₃	134456	2.017	A100
Ti ₀	75832	0.023	A100
Ti ₁	136760	0.151	A100
Ti ₂	141624	0.559	A100
Ti ₃	134456	1.291	A100
Os ₀	76346	0.011	V100
Os ₁	137274	0.042	V100
Os ₂	142138	0.126	V100
Os ₃	134970	0.289	V100
Os ₄	122234	0.462	V100
Os ₅	106974	0.666	V100

For the models at various values of ℓ_{max} , used to generate the results in Fig. 6. All NequIP models were trained using a single GPU. All training speeds were computed for training sets of 900 frames. The numerical coefficients of the irreducible feature representations were 64, 32, 16, 8, 4, and 2 for models with ℓ_{max} 5, 4, 3, 2, 1, and 0, respectively.

Finite size effects on learning difficulties

We also considered the effect of unit-cell size on the ability of these MLFFs to adequately explore atomic representations up to longer cutoff radii. This procedure is described in Section S7.B., where a 640 atom super-cell of Os was surveyed in AIMD at $1.25 \cdot T_{melt}$ and sequential frames were extracted to yield 200 train and 50 test frames. A coarse grid test was then performed using FLARE B2, up to an r_{cut} of 10 \AA , ℓ_{max} of 8, and n_{max} of 49, the results of which are provided in Fig. 7. These results demonstrate that even if more unique atomic environments are made available to the model radially, within the cutoff of the representation, the ‘best’ model still employs a short cutoff of 4 \AA , and high angular ($\ell_{max} = 6$ or 8) and radial resolution ($n_{max} = 9$ and larger). This answers the question posed from the frames sampled within the TM23 dataset, which are shorter ranged, such that most atomic environments within the representation cutoff are periodic images. These observations have direct implications in model design, where high angular and radial resolutions are required for early transition metals, which to this point, come with high computational cost to implement.

Influence of model accuracy on 0 K phonons

To illustrate the influence of model error on material properties, a subset of the FLARE B2 and NequIP models trained across TM23 were used to calculate phonon dispersions. Explicit details are provided in the ‘Methods’ Section. Results for Cu and Os are shown in Fig. 8, and exhibit different levels of accuracy, consistent with the trends discussed previously on energy, force, stress labels. Again, Cu represents an ‘easy’ metal, whereas Os is markedly more difficult to learn as evidenced by lower MLFF accuracy on forces, energies, and stresses, as well as lower learning exponents. Each model, using the full training set and the same as those presented in Fig. 2, is compared to ground-truth phonon dispersions obtained with DFT calculations. In Fig. 8, both NequIP and FLARE (using ACE B2 with $\zeta = 2$) models for Cu perform very well in predicting the phonon band structure compared to DFT, whereas this task is more difficult for Os, with

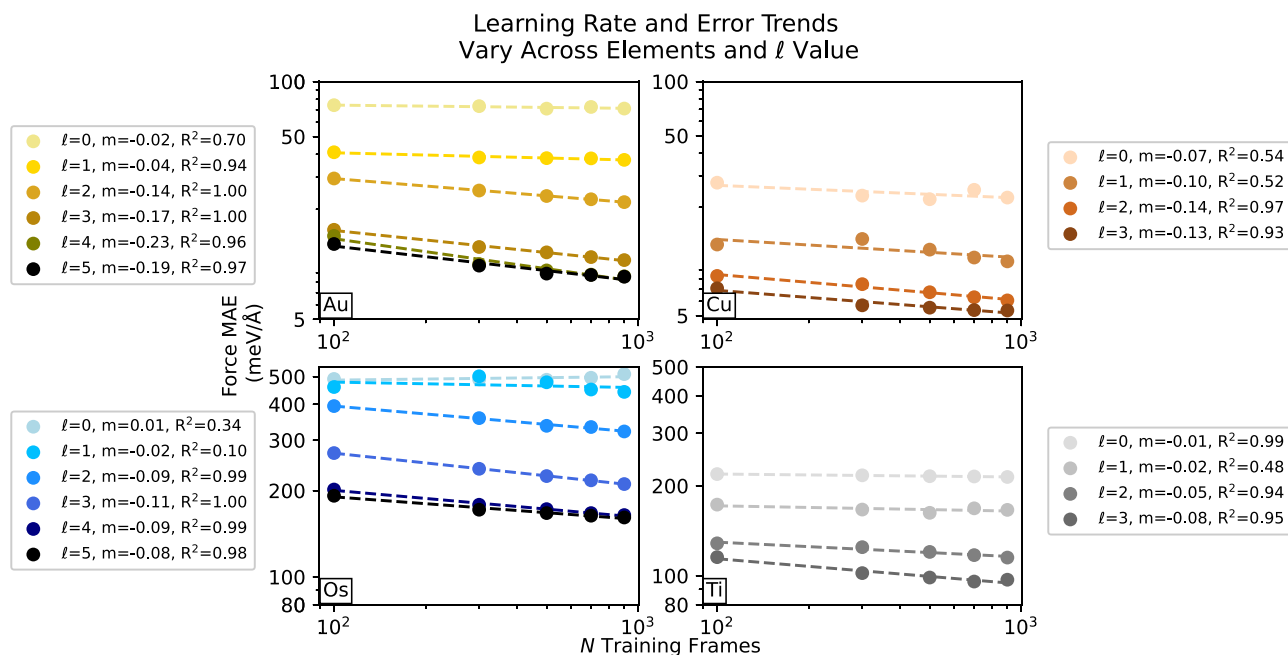


Fig. 6 | NequIP learning curves across ℓ_{max} . Model force MAE as a function of the number of training frames using varying ℓ_{max} values in NequIP for Au, Cu, Os, and Ti. Power-law fits are made to the data in log-log space to obtain the training exponents. To determine the robustness of each fit, R^2 values are also provided.

disagreement seen for various phonon band features. Acoustic phonon bands are better reproduced than optical phonon bands for NequIP and FLARE, with NequIP exhibiting better accuracy in both domains. On the other hand, acoustic and optical bands for Os are systematically underestimated by FLARE, but the overall shapes of the bands are in qualitative agreement with DFT. Systematic weakening of the phonon modes in this way can be explained by a slight overestimation of the Os HCP \hat{z} -lattice vector of the primitive unit-cell predicted by FLARE (4.51 Å versus 4.35 Å predicted by NequIP). A note must be made, however, that even though the NequIP model exhibits marginally more accurate vibrational spectra for both Cu and Os compared to FLARE using ACE B2, it is $\sim 200\times$ slower than FLARE. The trade-off between accuracy and efficiency can be made in the choice of models and architectures, and we provide the ‘best’ model parameters for both FLARE and NequIP in Section S6 from minimization of the force MAE (and maximization of the model likelihood for FLARE) to aide in this decision.

Periodic trends

The observation of systematic variation in MLFF accuracy across the composition space in the TM23 data set points to potential underlying physical and chemical differences in the interatomic interactions for different metals. By comparing the performance of atomic representations with different many-body order and angular resolution we can extract insights into the fundamental complexity and character of bonding in transition metals. To generate human intuition, we interpret these test error trends in terms of both chemical and structural properties of metals. Qualitatively, the test errors observed from Fig. 2 loosely follow a trend with respect to d -valence, where the metals with low numbers of valence d -electrons exhibit higher errors compared to metals with full d -shells. In addition to d -valence, the test errors also qualitatively follow a trend with respect to the initial crystal symmetry, as is shown in Supplementary Fig. 3. This is interesting, as the hexagonal crystals have two lattice parameters, which may provide a more difficult test-case with respect to the angular description of such materials. However, this does not explain the observed errors for the body-centered-cubic metals. We note that work from nearly 70 years ago exploring the relationship between electronic structure and crystal structure rationalizes the ground-state crystal symmetry attained by

metals to the differences in d -orbital participation in hybridization⁴⁸. In particular, they note that the weight of d -orbitals in hybridized orbitals attains a maximum near the middle of the d -block. Later, we will show a similar argument that helps to rationalize our error trends.

The higher relative errors of Group XII metals, disagreeing with this trend with d -valence, are explained by the test errors for forces, energies, and stresses of these metals being on the same order of magnitude as the DFT noise, which are then compared to low-magnitude MAVs for these metals with relatively low melting points. The trend with respect to d -valence was first evaluated in the context of several previous studies noting the appearance of directional-bonding ‘behaviors’ of early transition-metals^{49–51}. If substantiated, this correlation would partially explain why early transition metals require higher order angular resolution, as evidenced in the previous sections using both FLARE and NequIP. In refs. 49 and 51, the Cauchy pressure (defined by the relationship of elastic tensor components as $C_{12} - C_{44}$) is used as an indicator for directional bonding, which should be zero in pair-potentials. To determine if this trend was present, using Cauchy pressure as a potential label, we extracted available elastic constants for each metal in TM23 from the Materials Project repository and compared to the observed NequIP force % error. This comparison is provided in Supplementary Note 4. Ultimately, no correlation is observed, but the quality of this label to determine directional bonding has also come under question⁵².

Electronic structure and density of states influence MLFF accuracy

In this light, we additionally considered trends in another context, with respect to electronic structure, e.g. d -valence via d -band center, which is commonly used as a label to differentiate chemical behaviors among the transition metals (e.g. molecular adsorption and reactivity, especially in heterogeneous catalysis). The correlation of NequIP force % errors with the d -band center of the metals is provided in Supplementary Fig. 4, where a non-negligible linear trend is clearly observed between the force % test errors on the full set of data from NequIP and d -band centers directly computed here using a small set (3) of random frames from the melted test set of each metal. We also provide a complete representative set of the density of states plots of the TM23 metals in Supplementary Fig. 5. This

Fig. 7 | Large super-cell grid test over hyperparameters for Os using FLARE. Force % Error observed using FLARE B2, with power = 2 for a 640 atom unit cell of Os. The models are colored using a blue-white-red scale, with dark blue denoting lowest force % error. The best models employ $r_{cut} = 4.0 \text{ \AA}$, but with high values of ℓ_{max} of 6 and 8, and n_{max} of 9. Blank cells denote models that hit a memory limit using a single, 48 CPU node for FLARE.

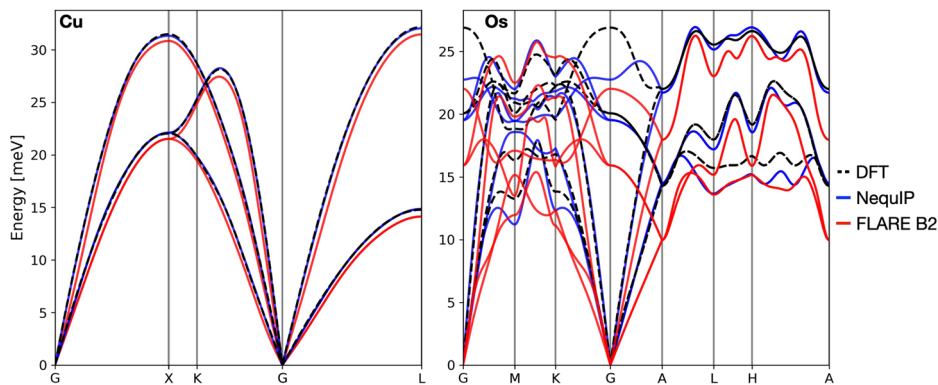
ℓ_{max}	n_{max}	Force % Error			
		cutoff			
		4.0	6.0	8.0	10.0
0	1	100.00	100.00	100.00	100.00
	5	21.37	34.92	46.15	59.70
	9	20.54	21.00	27.10	41.32
	13	20.14	19.98	19.47	26.58
	17	20.03	18.58	18.72	20.78
	21	20.00	18.05	18.21	19.17
	25	20.02	17.77	17.76	18.81
	29	20.04	17.70	17.49	
	33	20.10	17.68	17.38	
	37	20.22	17.70	17.17	
	41	20.34	17.78		
	45	20.46	17.87		
49	20.59	17.99			
2	1	57.59	72.00	82.44	89.77
	5	16.43	16.98	31.58	43.74
	9	15.92	15.08	19.63	32.09
	13	15.84	13.97	15.39	21.14
	17	15.99	13.08	14.28	16.82
	21	16.25	12.55	13.80	
	25	16.58	12.50		
	29	17.17	12.67		
	33	17.97	13.14		
	37	18.80			
41	19.81				
4	1	36.75	46.18	53.46	59.66
	5	9.36	11.41	15.63	25.06
	9	9.30	9.80	11.34	14.63
	13	9.40	9.07	11.66	
	17	9.82	8.66	10.07	
	21	10.67	8.72		
	25	11.85	9.10		
	29	13.31			
	33	14.73			
	37	16.43			
6	1	34.84	36.45	44.40	52.42
	5	8.38	10.56	11.68	13.18
	9	8.27	9.67	10.96	11.95
	13	8.39	9.24	10.11	
	17	8.80	8.92		
8	1	34.84	36.45	44.40	52.42
	5	8.38	10.56	11.68	13.18
	9	8.26	9.63	10.93	
	13	8.39	9.25		
	17	8.83	8.90		

result provides a strong indication that the discovered trends in test error across TM23 reveal a connection between the electronic structure and the many-body directional complexity of bonding present in these metals.

In seeking to quantitatively explain the test error trend among the transition metals with a measurable feature of the electronic structure, we also probed the correlation of test errors with DOS. This choice was made

because it gives us a lever to understand this relationship via the underlying DFT parameters, specifically how the smearing parameter σ_e within the VASP calculation influences the electronic occupations and resulting forces. Efforts by Drautz et al.^{53,54} also established differences in angular characters of transition metals, which required special treatment (e.g. explicit consideration of valence occupancy and moments of the density of states) for

Fig. 8 | Phonon dispersion curves for Cu and Os from DFT and MLFFs. DFT is a dashed, black curve, NequIP is blue, and FLARE B2 is red. Each model curve is from the full training set, and employs optimal model parameters determined from grid test.



the construction of bond order potentials. Methfessel and Paxton⁵⁵ showed that early transition metals have ‘more complicated’ density of states profiles, with attention paid to metals like Zr exhibiting steep variations in the DOS at E_{fermi} which, in the absence of smearing, results in ‘charge sloshing’ during SCF calculations as the Fermi level position fluctuates. We extend this logic to the complexity of the PES as a function of atomic positions and the resulting difficulty of learning the PES. Variations in local atomic configurations directly influence the electronic states and thus the energy of a configuration, and the occupancy changes of the bands (Kohn-Sham states at and near E_{fermi} are related to the potential energy surface). Given this relationship, we then hypothesized that the labels computed using DFT for early transition metals, which have sharper DOS at and near E_{fermi} , would thus be more sensitive to slight perturbations in the atomic positions than metals with smoother states about the Fermi level.

We tested this hypothesis of the complexity of the DOS by systematically smoothing out the occupations for fixed configurations of Au, Cu, Os, and Ti by recalculating the force/energy/stress labels with increasing values of electronic smearing σ_e . The intuition is that increasing the σ_e value smears the electron occupations, producing a smoother DOS and thus letting us explicitly test the relationship between the complexity of the DOS and the observed test errors (even letting σ_e vary to non-physical high values, e.g. 1.0 and 2.0 eV). Moreover, since we have direct access to the atomic prediction targets, we can also determine the effect of changing this smearing on the force magnitudes, and importantly, angular distributions of force direction. These results are provided in Fig. 9, where panel (a) contains test errors for FLARE models (with B2 and kernel power 2) trained on the recomputed labels with different values of artificial smearing. We can clearly see that the test error associated with models trained on increasingly smeared calculations decreases in Fig. 9a, with the high force % errors for Os and Ti drastically reducing, to such an extent that they resemble the errors of Au and Cu in the case of Ti (despite the dramatic differences in their electron count) when a value of $\sigma_e = 2.0$ eV is employed. While we acknowledge that these values of σ_e are extremely high, we find it striking that varying one parameter can so significantly reduce model fitting error with all else equal. This serves as a clue to what features of the elements are correlated with the regression difficulty.

In Fig. 9(b), we directly show the effect of increasing σ_e on the smoothness of the total d -DOS at and around E_{fermi} , with $\sigma_e = 2.0$ eV smoothing out all of the steep DOS features, most especially for Ti. We correlate this increase in smoothness, and thus decrease in complexity of the DOS with the sharp reduction in the test errors in panel (a). Moreover, we look at the differences of the radial and angular components of the force labels computed at $\sigma_e = 1.0$ and 2.0 eV relative to the original TM23 data at $\sigma_e = 0.2$ eV in panel (c). The middle-right column isolates any differences in the magnitudes of the forces for each atom across the entire training set, whereas the right-most column primarily isolates differences in the angular components, since the difference of the force vectors is first computed, then followed by the magnitude and mean. We note that the right-most column could also be shifted by a difference in the force magnitude. However, to

interpret these results, we have to look at the differences between the columns. If the values in the ‘angular’ (right-most) column are substantially larger than those in the ‘radial’ (middle-right) column, then we can quantitatively assert that the radial distribution of the force vectors is sensitive to the value of σ_e , and this change then results in the earlier transition metals becoming ‘learn-able’ using MLFFs with the same model architectures.

Perturbation theory explains relationship between density of states and angular sensitivity of forces

We explain the error trends further using arguments from perturbation theory. Consider a simple model system of an atom in a solid whose non-interacting electrons experience an external potential $V(r)$, which is produced by both the atom and the surrounding atoms in the solid. The electronic Hamiltonian is

$$\mathcal{H} = \frac{p^2}{2m} + V(r). \tag{3}$$

We assume for simplicity that this Hamiltonian produces eigenvalues and eigenvectors falling into one degenerate valence band (VB) and one degenerate conduction band (CB):

$$\text{VB} : \langle r|i \rangle = \phi(r)Y_L(r), \quad \mathcal{H}|i \rangle = 0 \tag{4}$$

$$\text{CB} : \langle r|j \rangle = \phi(r)Y_L(r), \quad \mathcal{H}|j \rangle = \Delta|j \rangle, \tag{5}$$

where $Y_L(r)$ are spherical harmonics, $L = \ell, m$ is a combined index for the principal and azimuthal quantum numbers, $\phi(r)$ is the radial component of the orbitals (with $\langle \phi(r)|\phi(r) \rangle = 1$), and Δ is the energy gap between the valence and conduction states. $\phi(r)$ is assumed to be the same for each orbital for simplicity. To match the state manifold of a typical transition metal, the combination of the VB and CB manifolds contains one s shell, one p shell, and one d shell, analogous to transition metal valence and conduction states. While these ‘bands’ do not account for bonding or the fact that the metal has no gap, they can be used as analogies to the manifold of states slightly above and below the Fermi level in a metal, with the key preserved feature being the different angular momentum character of each orbital. We assume the basic physical trends apply similarly well infinitesimally close to the Fermi level and a small but finite energy difference $\Delta/2$ away from the Fermi level.

The density of this system is

$$n(r) = \sum_{i \in \text{VB}} \langle r|i \rangle \langle i|r \rangle \tag{6}$$

$$n(r) = \sum_{i \in \text{VB}} |\phi(r)|^2 |Y_L(r)|^2 \tag{7}$$

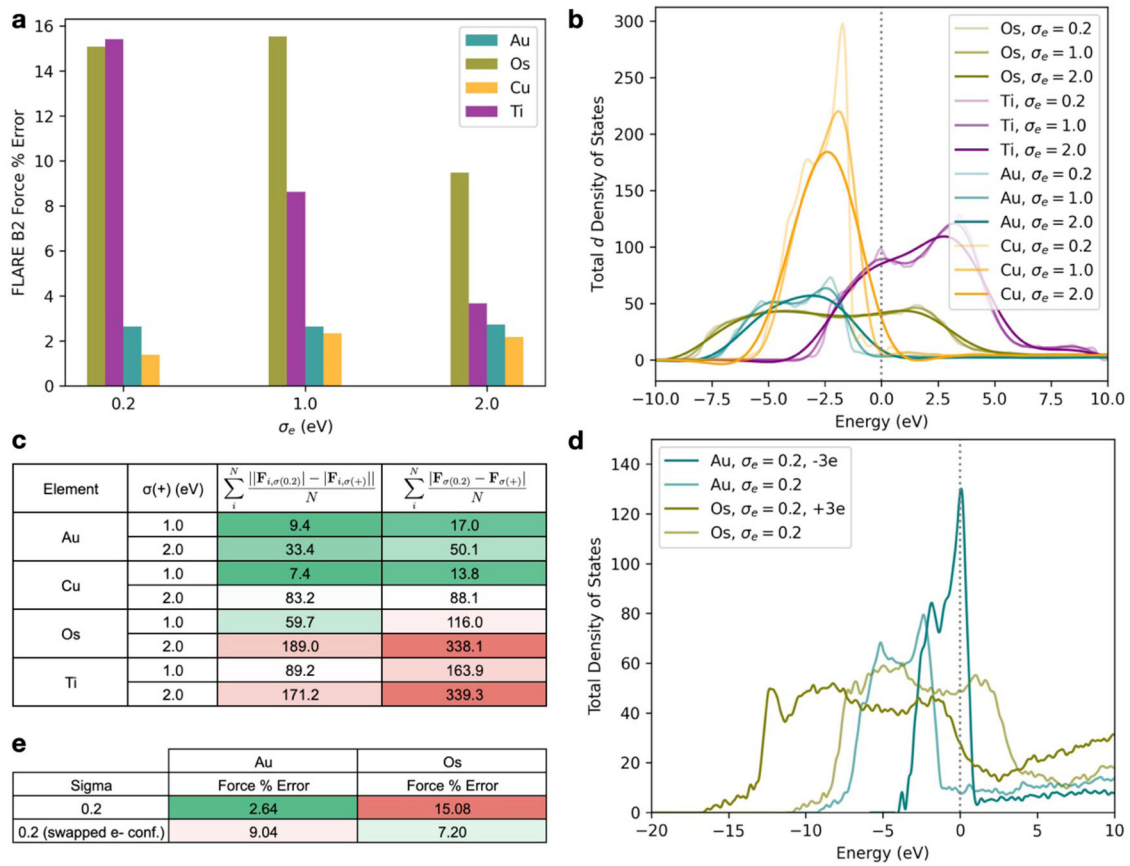


Fig. 9 | Force predictive accuracy is influenced with electronic structure of training data. **a** FLARE (B2 with kernel power = 2) force percent test errors on TM23 data recomputed using higher values of the electronic smearing σ_e for Au, Cu, Os, and Ti. **b** Total *d* density of states for the four metals with different values of σ_e . **c** Mean absolute differences between force magnitudes (left column) and mean of the magnitude of the difference between the force vectors. **d** Total density of states for

random frames from the ‘melt’ test-set for Au and Os with and without modified total numbers of electrons. The values provided in the legend represent the total change in the number of electrons per atom. **e** FLARE (B2 with kernel power = 2) force percent test errors for Au and Os with unmodified, and swapped electronic configurations to illustrate the effect of artificially moving the Fermi level relative to the *d*-DOS on the resulting model errors.

and the energy of the system is 0 (because the valence bands all have zero energy). Now consider a perturbing potential

$$V^{(1)}(r) = - \sum_L V_L Y_L(r). \quad (8)$$

This potential serves as a model for the effect of moving an ion in the solid, which changes the potential experienced by the electrons. By first order perturbation theory, the perturbation to the energy and valence orbitals is

$$|i^{(1)}\rangle = \sum_{j \in \text{CB}} -\frac{1}{\Delta} \langle j | V(r) | i \rangle | j \rangle \quad (9)$$

$$|i^{(1)}\rangle = \sum_L \sum_{j \in \text{CB}} \frac{C_{L_i L_j L} V_L}{\Delta} |j\rangle, \quad (10)$$

where $C_{L_i L_j L} = \int d\Omega Y_{L_i}(\Omega) Y_{L_j}^*(\Omega) Y_L(\Omega)$ (same as the Gaunt coefficients except for the complex conjugation). The resulting perturbation to the density is

$$n^{(1)}(r) = \frac{|\phi(r)|^2}{\Delta} \sum_{LL'} Y_L(r) V_{L'} P_{LL'} + \text{C.C.} \quad (11)$$

$$P_{LL'} = \sum_{i \in \text{VB}} \sum_{j \in \text{CB}} C_{L_i L_j L}^* C_{L_i L_j L'}, \quad (12)$$

where C.C. stands for the complex conjugate of the expression. With non-integer occupation numbers f_b , the density response $P_{LL'}$ becomes

$$P_{LL'} = \sum_{i \in \text{VB}} \sum_{j \in \text{CB}} (f_i - f_j) C_{L_i L_j L}^* C_{L_i L_j L'}. \quad (13)$$

Similarly, the energy perturbation to first order is

$$E^{(1)} = - \sum_L V_L \left(\sum_{i \in \text{VB}} f_i C_{L_i L_i L} + \sum_{j \in \text{CB}} f_j C_{L_j L_j L} \right). \quad (14)$$

Because the forces on the nuclei are determined by the electronic charge distribution, we can interpret the trends in the MLFF force errors by analyzing the density response to a potential given by Eq. (11), and in particular the angular dependence of the response given by Eq. (13). In general, we can expect that higher-order spherical harmonics in the density response will result in the forces being more difficult to predict, since the forces are determined by the interaction between the ions and electron density distribution. Because $C_{L_1 L_2 L_3} = 0$ if $\ell_1 + \ell_2 < \ell_3$, *d* to *d* transitions both respond to higher-order V_L terms (up to $\ell = 4$) and produce higher-order density fluctuations in $n^{(1)}(r)$ via Equation (11). Therefore, we focus on the

role of these d to d transitions for different metals to explain the complexity of the PES.

In the limiting case where the d shell is mostly occupied (i.e. the late transition metals in Groups X–XII), there are few or no conduction d electrons, so the sum over j in Equation (13) vanishes, resulting in a less complex density response and therefore a simpler PES. However, when there are d orbitals in both the conduction and valence bands, the sum over j has finite contributions, and $\ell = 4$ responses to the potential arise, making the PES more complicated. In this simplified perturbation theory model, the magnitude of the $\ell = 4$ density response should be roughly proportional to $(10 - N_d)N_d$, where N_d is the number of d electrons per atom in the system. This intuition is confirmed for the metals with full d shells (Groups X–XII) and the roughly half-filled shells (Groups VI–IX), as the latter have much higher force errors than the former (Fig. 2). We also show this to be true empirically, by artificially changing the number of electrons in the Au and Os systems, effectively swapping their electronic configurations, which changes the relative position of the Fermi level to the DOS, shown in Fig. 9d, and recomputing the FLARE force percent errors on the test set in Fig. 9e. We see that by changing the number of electrons, and thus shifting the location of the Fermi energy, we can directly influence the difficulty of learning the force labels.

In spite of having fewer d electrons than the metals with half-filled shells, the metals in Groups IV and V with mostly empty (but importantly, not completely empty) shells also have high force errors. We attribute this to the fact that these metals still have a large d -character DOS both above and below the Fermi level, similarly to the middle-group transition metals, whereas the later-group transition metals have very little unoccupied d -DOS near the Fermi level. This trend is illustrated in Supplementary Fig. 5, which shows that transition metals in Groups IV–IX have a large d -DOS both above and below the Fermi level, whereas the d -DOS for Groups X–XII is mostly or entirely below the Fermi level. Therefore, we expect the early transition metals to each have a complex PES like the middle transition metals, which is confirmed by Fig. 2.

The effects of smearing can also be considered within the perturbation theory model. Empirically, large, non-physical values of smearing significantly decrease the force errors for early and middle transition metals (Fig. 9a). The case of the earlier transition metals is complicated, but there are a few reasons to expect the smearing to decrease the force errors. First, smearing out the occupations causes all the m channels of the d shell to become partly occupied (i.e. $f_i \approx f_j$), which damps the density response in Eq. (13). Second, as shown in Fig. 9b, smearing evens out the DOS of the individual d orbitals, which results in similar occupation numbers f_i for different m values, which also causes Eq. (13) to vanish. Finally, larger smearing smooths out the total d DOS near the Fermi level, especially in difficult transition metals like Ti and Os (Fig. 9b), which might also smooth the PES.

Ultimately, early transition metals are found to possess more structurally sensitive many-body interactions, which we find to be correlated with their more complicated electronic structures, and thus require higher angular resolution of the MLFF tasked with learning forces, energies, and stresses. Confirming this is also the overall reduction in test error magnitude as the angular resolution of NequIP is increased, as was demonstrated in Section “TM23 Data Set Description”. This finding motivates further equivariant MLFF development to access higher values of ℓ_{max} with increased computational efficiency, and will be the followup investigation to this work as these novel MLFF methods have recently become available (e.g. Allegro⁴² and MACE³⁷).

Discussion

This work presents TM23, a benchmark data set for transition metals comprised of high-fidelity first-principles calculations, and provides a systematic comparison of two leading MLFF architectures. FLARE and NequIP models were tasked with learning force, energy and stress labels from the same reference data, and marked differences in performance were observed. These findings uncover persistent trends in MLFF accuracy across TM23,

with interatomic interactions in early transition metals (Group VIII and below) being demonstrably more difficult to capture accurately than Group IX, X, XI, and XII metals. These trends were observed not only in final model accuracy but also in learning exponents, as a function of training set sizes using NequIP. Importantly, these trends were then explained using a detailed understanding of the electronic structures of these metals coupled with an explanation rooted in perturbation theory. This work highlights the utility of using systematically controllable geometric representations, such as the atomic cluster expansion and equivariant neural networks, to uncover subtle and complex physical features of metallic interactions in a data-driven way from high-fidelity first-principles reference data. The accuracy trends for FLARE and NequIP across the space of transition metals point to the varying degree of importance of directionality and many-body character of interactions, the complexities of which can be artificially reduced via modification of the electronic structure in the underlying quantum mechanical calculation. Ultimately, the TM23 data set provides some of the most challenging reference benchmarks for currently leading MLFF approaches, even if just for single-element bulk systems. Substantial improvements are therefore still needed and motivated by these results for future MLFF architectures, focusing on more efficient many-body representations that can expand to higher radial and angular resolutions without sacrificing computational efficiency. We prove this empirically, by establishing the sensitivity of the radial distribution of the forces to the electronic bonding interactions within the metal. Systematic work is needed on using this challenging data set to benchmark different MLFF model architectures and regression methods. Another extension of this work can target the augmentation of this data set, from which purpose-built models can be trained for property prediction (e.g. melting point estimation, dislocation dynamics, etc). We anticipate that our findings and reference data will help both to anticipate appropriate model parameters for practitioners studying transition metals and to advance model development for atomistic description of heterogeneous metal catalysts, multi-element alloys and many other applications where these elements are present.

Methods

Training data acquisition: density-functional theory and ab initio molecular dynamics

The complete computational workflow is detailed in Fig. 1. Bulk configurations of all metals were extracted from the Materials Project (MP) repository¹⁶, where the lowest-energy crystalline phase was selected for each system. This workflow was employed for all systems, including those exhibiting dynamic instabilities⁵⁶ in bcc phases at lower temperature (e.g. Ti, Zr, and Hf). Each phase was selected based on the convex hull energy, the lowest being chosen. The selected phase was also confirmed to be experimentally observed, as is provided on the MP ‘Materials-Explorer’ dashboard. Reference MP identification numbers (MP-IDs) corresponding to each metal structure employed are provided in Supplementary Table 1. Super-cells of each metal were then created, sized such that at least 7 Å of spacing existed between an atom and its periodic images. This lattice requirement was chosen as a lower bound since previous FLARE force field training displayed sufficient accuracy in modeling 2-body behavior at this distance^{10,43}. Computational efficiency in creating the TM23 data set was also considered, as super-cells of this size contained 32 to 71 atoms. The pymatgen library was employed to simplify super-cell creation and calculation setup⁵⁷. To increase the diversity of atomic environments within each super-cell, a single vacancy defect was introduced by randomly deleting an atom. Intuitively, introduction of a vacancy should increase the number of nontrivial atomic environments from the onset of each trajectory, and may facilitate more complicated dynamics (e.g., vacancy diffusion) throughout the course of the simulation.

Following creation of each super-cell and incorporation of a single vacancy, ab initio molecular dynamics (AIMD) simulations were then performed in three thermal regimes: ‘cold’ metals surveyed at 25% of their experimental melting temperatures T_{melt} , ‘warm’ metals at $0.75 \cdot T_{melt}$, and ‘melt’ molten configurations from dynamics surveyed at $1.25 \cdot T_{melt}$. These

relative temperatures were chosen such that diverse atomic orderings would be captured within the training set, ranging from fully crystalline to amorphous atomic environments. The experimental melting temperatures associated with all elements are listed in Supplementary Table 2.

AIMD trajectories and density functional theory (DFT) calculations were completed using the Vienna Ab initio Simulation Package (VASP, version 5.4.1)^{58–61}. The pseudopotentials employed for each metal are listed in Supplementary Table 2. AIMD trajectories were surveyed for a total of 55 picoseconds using a timestep of 5 femtoseconds for all metals. The NVT ensemble with the Nosé-Hoover thermostat^{62,63} and a Nosé-mass of 40 timesteps were employed. To reduce the computational cost incurred by each AIMD trajectory, since sequential configurations of successive frames are strongly correlated, eigenvalues for the wavefunction were only sampled using the Γ k-point. While the density of the k-point mesh could influence the dynamics that are observed throughout the course of a trajectory, one can rely on this effect being less pronounced at longer timescales. This assumption is valid given the stochastic nature of MD trajectories, especially using a canonical ensemble like NVT with the Nose-Hoover thermostat.

An important discussion is provided here with respect to the exclusion of spin-polarization from both the AIMD trajectories and high-fidelity DFT calculations. Spin-polarization can have a pronounced effect on the electronic structure for magnetic systems, and has been shown to influence the vacancy diffusion energy⁶⁴ up to 300 K. Moreover, we note that property predictions could be affected by magnetic effects, and use of these force fields would require caution under such conditions. However, all metals in both AIMD trajectories at the ‘warm’ and ‘melt’ temperatures are well above their Curie temperatures, so in ‘real-world’ dynamics at these elevated temperatures, spin ordering in the electronic degrees of freedom would not be present. Thus, there is an inherent issue with combining high-temperature AIMD for dynamics and static DFT for energy, force, and stress label calculation: even if the configuration itself has a high potential energy (i.e. it is far from an equilibrium 0 K ordered structure) and was sampled from an AIMD trajectory above the Curie temperature, the configurations are ultimately endowed with labels calculated for a wavefunction solved at approximately 0 K electron temperature. Finite temperature electronic smearing is supplied to each static DFT calculation via σ_e in VASP, serving as a crude approximation to elevated electron temperature. The σ_e value in VASP was initially not changed to approximate the electronic temperature of the systems as the temperature of the AIMD simulations was varied. However, this value was varied to even non-physically high values of 1.0 and 2.0 eV for a subset of the metals to provide the results in Fig. 9 of the main text. Mixing the ‘warm’ and ‘melt’ training labels with ‘cold’ metals below the Curie temperature, without an MLFF designed to account for the inclusion of spin would likely cause the model to try to learn two different potential energy surfaces— one with spin ordering and one without. Therefore, we compromised by neglecting spin entirely for both AIMD trajectories and the subsequent high-fidelity DFT calculations. While this means that the sampled configurations and computed training labels for metals with significant spin interactions (e.g. Fe, Co, Ni) are cruder approximations to their ‘real’ dynamics and potential energy surfaces, we leave these questions to be answered in a future investigation.

Individual frames were then extracted from each AIMD trajectory at intervals of 50 fs, excluding the first 5 ps of the trajectory. Frames from the first 5 ps were excluded so that the extracted frames were given sufficient time to equilibrate with the applied thermostat. This selection procedure resulted in 1000 frames from each trajectory, yielding a total of 3000 frames for each metal. In order to visualize the extent of the thermal disorder in our trajectories, radial distribution functions averaged across all trajectories were computed using the ASE package. The methods and RDF plots are provided in the Supplementary Information.

Higher-fidelity single point energy, force, and stress labels were then calculated using increased k-mesh densities converged on a per-element basis. The k-point spacing was chosen such that the energy noise per atom was below 1 meV per atom and the force noise was below 5 meV \AA^{-1} . Element specific k-point grids were employed to respect inherent

symmetries of each system, as recommended by VASP. To further facilitate convergence, Methfessel-Paxton smearing at the Fermi-level was employed, with all metals using a value of 0.2 eV, but was eventually varied given to generate the labels corresponding to the models trained in Fig. 9. Moreover, Fig. 9(d) employed the VASP ‘NELECT’ parameter, which allowed us to artificially change the total number of electrons for the Au and Os labels, effectively testing the effect of swapping their electronic configurations. All calculations employed a cutoff energy of 520 eV, which was sufficient for all ENMAX values provided by the pseudopotentials for all metals.

Element specific semi-core corrections and k-point densities

In the interest of reproducibility, we also provide a complete description of the employed semi-core corrections used in each pseudopotential across the metals, as well as the minimum k-point densities applied for each system, provided in Supplementary Table 1. Moreover, to address the pertinent question of DFT accuracy of this data set, we evaluate the correlation between NequIP predicted force, energy, and stress percent errors (using the full training set) against the minimum k-point density observed along the lattice vectors of each cell in Supplementary Fig. 1. The pseudopotential naming convention maintains consistency with those presented in the VASP documentation.

Gaussian processes in the fast learning of atomistic rare events (FLARE) architecture

The Gaussian process machine-learning architecture implemented in FLARE incorporates the atomic cluster expansion (ACE) descriptors and the normalized dot-product kernel, described in detail elsewhere⁸. Unlike the 2+3 body atomic environment representations used in previous implementations of FLARE¹⁰, the representations for each individual atomic environment in the current iteration are computed using the atomic cluster expansion (ACE) of Drautz³⁶. Briefly, ACE represents the local environment around each atom using a ‘fingerprint’ which projects the distribution of neighboring atoms into a set of radial and angular basis functions. The GP then compares the full set of atomic environments in the test frame to other descriptor vectors in the training set to perform inference, which provides the predictive energy, forces and stress, as well as quantitative uncertainties. Here, we maintain consistency with the notation of refs. 8 and 36. We use B1 and B2 descriptors from the equation (28) and (29) of the original ACE paper³⁶, corresponding to rotation-invariant 2-body and 3-body representations. We note that using descriptors of higher body order can improve the accuracy, but is computation and memory intensive, especially considering the construction of SGP requires storage of the descriptors and their gradients of the entire training data set. To develop a more scalable method for higher body order descriptors and benchmark their performance with the SGP will be the focus of our future work. For the kernel of SGP, we choose the normalized dot product raised to the power (ζ) of 1 and 2, which lifted the body order of our model. Specifically, the 2-body B1 descriptors with $\zeta = 2$ becomes effectively 3-body, and the 3-body B2 descriptors with $\zeta = 2$ becomes effectively 5-body. We provide ‘best’ model parameters for both kernel powers, but only results for $\zeta = 2$, since these are shown to be systematically more accurate across all labels and metals.

Our GP model is trained by optimizing hyperparameters to maximize the log likelihood function which describes the overall agreement of the model with training data and the complexity of the model. For each model, the likelihood is optimized via gradient descent with respect to four hyperparameters: the signal variance (σ), and the three noise variances for each of energy, force, and stress (σ_E , σ_F , and σ_S , respectively). The initial value of signal variance is set to 2.0. Each noise parameter is initialized to the expected error for the corresponding quantity, specifically: 0.001 eV per atom, 0.05 eV \AA^{-1} , and 0.005 eV \AA^{-3} , respectively. Additionally, depending upon the temperature of the AIMD data fed into the model, the initial σ_F is varied accordingly, since the force MAE scales with the temperature of the AIMD trajectory (higher temperature configurations tend to have larger forces acting on the atoms). For all FLARE training, hyperparameter optimization was completed periodically throughout training until the final

frame addition, which varied depending on the size of the training set (from a total of 100 to 2700 frames), as hyperparameter optimization requires re-computing and inverting the covariance matrix at each step. Furthermore, the gradient tolerance for the convergence criteria of L-BFGS-B optimization method was set to $1\text{E-}4$ for the marginal log likelihood, and the maximum number of iterations was set to an appreciably large value (200) to ensure convergence, if necessary.

Lastly, an exhaustive grid-search through reasonable values of the descriptor parameters was also completed for each metal and B1 and B2 respectively. These descriptor parameters are not optimized against the likelihood at training time like conventional hyperparameters, and can strongly influence model behavior. The radial cutoff r_{cut} , radial basis length (n_{max}), and angular basis length (l_{max}) were all tested over a broad range of reasonable values. Interested readers may find the full results in Tables S13–S16, whereas the main text only presents results for models that employ the best model parameters for each ACE descriptor, which were selected using a combination of maximum likelihood and minimum force MAE. If a system presented the scenario where force MAE was minimized with different parameters when compared to maximum likelihood, the parameters were taken from the minimum force MAE calculation. This scenario is plausible since FLARE training was done on forces, energies, and stresses.

Equivariant message-passing graph neural network (NequIP) architecture

We also employ the equivariant message-passing interatomic potential NequIP¹¹. NequIP has recently been shown to be highly sample-efficient, and to be remarkably robust when compared to other existing ML methods in the MLFF literature on main-group benchmarks. The NequIP architecture is described in detail elsewhere¹¹, but the key idea lies in learning featurizations of the atomistic structure which are explicitly constructed to be equivariant under symmetry operations of the Euclidean group $E(3)$. $E(3)$ is comprised of translations, rotations, and mirror operations, covering the physical symmetries present in atomistic systems. Equivariance is distinct from invariance by the fact that invariant quantities do not vary under $E(3)$ operations, whereas equivariant quantities transform appropriately with $E(3)$ operations. In other words, an invariant scalar does not change under symmetry transformations, whereas e.g. an equivariant vector transforms in a way that is commensurate with the group (see equation 3 of ref. 42). An example of an invariant MLFF is the SchNet potential⁴⁰, which only operates on invariant descriptions of the geometry (distances r_{ij}), whereas the equivariant NequIP potential additionally uses higher order tensor representations to encode more complex geometric information about atomic environments. NequIP uses these features in a message-passing framework that represents the atomistic structure as a graph, and makes predictions by iteratively propagating information along that graph through a series of N_{layer} update layers.

The training procedure for NequIP is similar to that of FLARE, where the train-test splits are equivalent, but NequIP additionally withholds a further percentage of the training data, 10% in this work, that is used to monitor the progress of the ongoing training procedure. As was done for FLARE, an exhaustive grid search over the hyperparameters of the model was completed for each metal. The hyperparameters scanned for NequIP were the: radial cutoff (r_{cut}), number of layers (N_{layers}), angular resolution of the network (l_{max}), number of features ($f_{irreducible}$), learning rate, and force/energy weights. The best model parameters for each metal are provided in Supplementary Note 6. The best model for each metal was chosen via minimization of the force error on the validation set. To limit computational inefficiency in observing convergence for each NequIP model, a learning rate scheduler was employed, where the learning rate was reduced by 80% if the force MAE on the validation set did not improve within 100 epochs, and training is concluded if the learning rate reduced to be less than $1\text{E-}05$. In the loss function, energies were weighted using the per-atom-MSE, and the loss coefficients were set to 1 for both the forces and stresses, shown in Eqn. 30 of ref. 42. To limit the amount of computational resources required for this component, we set a hard-limit of 3 weeks of wall-time for training on a

single A100 GPU. Several models converged before this limit, but most were halted at this upper bound.

Phonon dispersion

Phonon dispersion curves were calculated for Cu and Os using the Phonopy⁶⁵ and Phoebe packages⁶⁶. First, the $(1 \times 1 \times 1)$ primitive unit-cell of each material was relaxed using LAMMPS, isotropically along each lattice vector using conjugate gradient descent until energy and force thresholds of 1×10^{-12} were met. To perform phonon calculations, Phonopy was used to generate displaced super-cell structures from which forces for each super-cell were calculated using FLARE and NequIP potentials. Then, Phonopy was used to collect the force constants from each super-cell calculation and Phoebe was applied to construct the dynamical matrix and plot the phonon dispersions.

DFT phonon calculations were then completed for each material using the same workflow in order to provide ‘ground-truth’ labels with which to compare the MLFF predictions. These DFT calculations were performed using the same VASP pseudopotentials, k -point densities, and INCAR parameters as the TM23 frames for Cu and Os.

Data availability

The data and related code in this paper are published in Materials Cloud Archive⁶⁸ with <https://doi.org/10.24435/materialscloud:6c-b3>⁶⁹.

Code availability

The details about VASP, a proprietary code, can be found at <https://www.vasp.at/>. The details about FLARE, and NequIP, which are open-source codes can be found at <https://github.com/mir-group/flare> and <https://github.com/mir-group/nequip>, respectively.

Received: 20 March 2023; Accepted: 10 April 2024;

Published online: 07 May 2024

References

- Mendelev, M. I. et al. Development of new interatomic potentials appropriate for crystalline and liquid iron. *Philos. Mag.* **83**, 3977–3994 (2003).
- Daw, M. S. & Baskes, M. I. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Phys. Rev. B* **29**, 6443–6453 (1984).
- Baskes, M. I. Application of the embedded-atom method to covalent materials: A semiempirical potential for silicon. *Phys. Rev. Lett.* **59**, 2666–2669 (1987).
- Finnis, M. W. & Sinclair, J. E. A simple empirical n-body potential for transition metals. *Philos. Mag. A* **50**, 45–55 (1984).
- Van Duin, A. C., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: a reactive force field for hydrocarbons. *J. Phys. Chem. A* **105**, 9396–9409 (2001).
- Chenoweth, K., Van Duin, A. C. & Goddard, W. A. ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *J. Phys. Chem. A* **112**, 1040–1053 (2008).
- Senftle, T. P. et al. The ReaxFF reactive force-field: development, applications and future directions. *npj Comput. Mater.* **2016** *2:1* **2**, 1–14 (2016).
- Vandermause, J., Xie, Y., Lim, J. S., Owen, C. J. & Kozinsky, B. Active learning of reactive bayesian force fields applied to heterogeneous catalysis dynamics of h/pt. *Nat. Commun.* **13**, 5183 (2022).
- Johansson, A. et al. Micron-scale heterogeneous catalysis with bayesian force fields from first principles and active learning. *Preprint at* <https://arxiv.org/abs/2204.12573> (2022).
- Vandermause, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 1–11 (2020).
- Batzner, S. et al. E(3)-Equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022** *13:1* **13**, 1–11 (2021).

12. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203 (2017).
13. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
14. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014 1:1** 1, 1–7 (2014).
15. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 248–255 (2009).
16. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
17. Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
18. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
19. Byggmästar, J., Nordlund, K. & Djurabekova, F. Gaussian approximation potentials for body-centered-cubic transition metals. *Phys. Rev. Mater.* **4**, 093802 (2020).
20. Rosenbrock, C. W. et al. Machine-learned interatomic potentials for alloys and alloy phase diagrams. *npj Comput. Mater.* **7**, 24 (2021).
21. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
22. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
23. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
24. Christensen, A. S. & von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn.: Sci. Technol.* **1**, 045018 (2020).
25. Blank, T. B., Brown, S. D., Calhoun, A. W. & Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **103**, 4129–4137 (1995).
26. Handley, C. M., Hawe, G. I., Kell, D. B. & Popelier, P. L. Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning. *Phys. Chem. Chem. Phys.* **11**, 6365–6376 (2009).
27. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
28. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
29. Xie, Y., Vandermause, J., Sun, L., Cepellotti, A. & Kozinsky, B. Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene. *npj Comput. Mater.* **7**, 1–10 (2021).
30. van der Oord, C., Sachs, M., Kovács, D. P., Ortner, C. & Csányi, G. Hyperactive learning for data-driven interatomic potentials. *npj Comput. Mater.* **9**, 168 (2023).
31. Zhang, L., Lin, D.-Y., Wang, H., Car, R. & E. W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
32. Xie, Y. et al. Uncertainty-aware molecular dynamics from bayesian active learning for phase transformations and thermal transport in sic. *npj Comput. Mater.* **9**, 36 (2023).
33. Glielmo, A., Zeni, C. & De Vita, A. Efficient nonparametric N-body force fields from machine learning. *Phys. Rev. B* **97**, 184307 (2018).
34. Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Sim.* **14**, 1153–1173 (2016).
35. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **115**, 1051–1057 (2015).
36. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
37. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In Koyejo, S. et al. (eds.) *Adv. Neural Inf. Process. Syst.*, vol. 35, 11423–11436 (Curran Associates, Inc., 2022). https://proceedings.neurips.cc/paper_files/paper/2022/file/4a36c3c51af11ed9f34615b81edb5bbc-Paper-Conference.pdf.
38. Schütt, K., Unke, O. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Int. Conf. Mach. Learn.*, 9377–9388 (PMLR, 2021).
39. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
40. Schütt, K. et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Guyon, I. et al. (eds.) *Adv. Neural Inf. Process. Syst.*, vol. 30 (Curran Associates, Inc., 2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf.
41. Wang, H., Zhang, L., Han, J. & E. W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
42. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
43. Lim, J. S. et al. Evolution of Metastable Structures at Bimetallic Surfaces from Microscopy and Machine-Learning Molecular Dynamics. *J. Am. Chem. Soc.* **142**, 15907–15916 (2020).
44. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
45. James, A. M. & Lord, M. P. *Macmillan's chemical and physical data* (Macmillan, London, 1992).
46. Kovács, D. P. et al. Linear atomic cluster expansion force fields for organic molecules: beyond RMSE. *J. Chem. Theory Comput.* **17**, 7696–7711 (2021).
47. Cheng, B., Engel, E. A., Behler, J., Dellago, C. & Ceriotti, M. Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci.* **116**, 1110–1115 (2019).
48. Altmann, S. L., Coulson, C. A. & Hume-Rothery, W. On the relation between bond hybrids and the metallic structures. *Proc. R. Soc., Lond., Ser. A. Math. Phys. Sci.* **240**, 145–159 (1957).
49. Eberhart, M. E. & Jones, T. E. Cauchy pressure and the generalized bonding model for nonmagnetic bcc transition metals. *Phys. Rev. B* **86**, 134106 (2012).
50. Mrovec, M., Nguyen-Manh, D., Elsässer, C. & Gumbsch, P. Magnetic bond-order potential for iron. *Phys. Rev. Lett.* **106**, 246402 (2011).
51. Pettifor, D. G., Aoki, M., Murrell, J. N., Cottrell, A. & Stoneham, A. M. Bonding and structure of intermetallics: A new bond order potential [and discussion]. *Philos. Trans. A Math. Phys. Eng. Sci.* **334**, 439–449 (1991).
52. Gong, X. et al. Exploring alloying effect on phase stability and mechanical properties of γ -ni3nb precipitates with first-principles calculations. *Mater. Des.* **196**, 109174 (2020).
53. Drautz, R., Hammerschmidt, T., Čák, M. & Pettifor, D. G. Bond-order potentials: derivation and parameterization for refractory elements. *Model. Simul. Mat. Sci. Eng.* **23**, 074004 (2015).
54. Drautz, R. & Pettifor, D. G. Valence-dependent analytic bond-order potential for transition metals. *Phys. Rev. B* **74**, 174117 (2006).
55. Methfessel, M. & Paxton, A. T. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B* **40**, 3616–3621 (1989).
56. Qian, X. & Yang, R. Temperature effect on the phonon dispersion stability of zirconium by machine learning driven atomistic simulations. *Phys. Rev. B* **98**, 224108 (2018).

57. Ong, S. P. et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
58. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
59. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B - Condens. Matter Mater. Phys.* **54**, 11169–11186 (1996).
60. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
61. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B - Condens. Matter Mater. Phys.* **59**, 1758–1775 (1999).
62. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **31**, 1695 (1985).
63. Martyna, G. J., Klein, M. L. & Tuckerman, M. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635 (1998).
64. Angsten, T., Mayeshiba, T., Wu, H. & Morgan, D. Elemental vacancy diffusion database from high-throughput first-principles calculations for fcc and hcp structures. *New J. Phys.* **16**, 015018 (2014).
65. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
66. Cepellotti, A., Coulter, J., Johansson, A., Fedorova, N. S. & Kozinsky, B. Phoebe: A high-performance framework for solving phonon and electron boltzmann transport equations. *J. Phys.: Mater.* **5**, 035003 (2022).
67. Torrisi, S. B. J. *Materials Informatics for Catalyst Stability & Functionality*. PhD thesis, Harvard University, Cambridge, MA, USA (2021). Available at https://nrs.harvard.edu/URN-3:HUL_INSTREPOS:37368388.
68. Talirz, L. et al. Materials cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
69. Owen, C. J. et al. *Materials Cloud Archive* <https://doi.org/10.24435/materialscloud:6c-b3> (2024).

Acknowledgements

We gratefully acknowledge Bill Curtain, Georg Kresse, Ralf Drautz, Gabor Csanyi, and Karsten Jacobsen for thoughtful discussions regarding interpretation of the data, methods, and training protocols. We thank Dr. Jin Soo (David) Lim and Dr. Jonathan Vandermause for helpful conversations at the outset of this project. This work was supported by the US Department of Energy, Office of Basic Energy Sciences Award No. DE-SC0022199 and No. DE-SC0020128, as well as by Robert Bosch LLC. C.J.O. is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. S.B.T. was supported by the Department of Energy Computational Science Graduate Fellowship under grant DE-FG02-97ER25308. Computational resources were provided by the FAS Division of Science Research Computing Group at Harvard University.

Author contributions

C.J.O. created the data set, compiled and analyzed the data, performed all model training and evaluation using FLARE and NequIP architectures, and co-wrote the manuscript. S.B.T. initiated the project, built Python scripts for data management and calculations, and co-wrote the manuscript. Preliminary results from an early version of work were published in S.B.T.'s doctoral dissertation⁶⁷. Note that all data presented here is different from those presented in the dissertation. S.B.T. and C.J.O. jointly designed the figures. Y.X. aided in FLARE implementation and calculations, and provided detailed feedback on the FLARE methods section. S.B. and A.M. provided discussions regarding the NequIP architecture and feedback for the respective methods section. J.C. provided helpful discussion and analysis of the phonon dispersion results obtained via Phonopy and Phoebe codes and wrote the corresponding methods section. K.B. provided helpful discussions pertinent to the final interpretation of the error trends, and was responsible for the perturbation theory considered in this context. L.S. advised components of this work. B.K. supervised all aspects of this work. All authors contributed to revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-024-01264-z>.

Correspondence and requests for materials should be addressed to Cameron J. Owen or Boris Kozinsky.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024